Effects of kernel function on nu support vector machines in extreme cases

AUTHOR(S):

Ikeda, K

CITATION:

Ikeda, K. Effects of kernel function on nu support vector machines in extreme cases. IEEE TRANSACTIONS ON NEURAL NETWORKS 2006, 17(1): 1-9

ISSUE DATE:

2006-01

URL:

http://hdl.handle.net/2433/50322

Effects of Kernel Function on Nu Support Vector Machines in Extreme Cases

Kazushi Ikeda, *Member, IEEE*

***Abstract*—How we should choose a kernel function in support vector machines (SVMs), is an important but difficult problem. In this paper, we discuss the properties of the solution of the ν-SVM's, a variation of SVM's, for normalized feature vectors in two extreme cases: All feature vectors are almost orthogonal and all feature vectors are almost the same. In the former case, the solution of the ν-SVM is nearly the center of gravity of the examples given while the solution is approximated to that of the ν-SVM with the linear kernel in the latter case. Although extreme kernels are not employed in practice, analyzes are helpful to understand the effects of a kernel function on the generalization performance.**

***Index Terms*—Asymptotic properties, generalization ability, kernel method, ν-SVM, support vector machine (SVM).**

I. Introduction

In a decade, support vector machines (SVMs) have attracted much attention as a new classification technique with good generalization ability [1]–[5]. The basic idea of SVMs is to map input vectors into a high-dimensional feature space and linearly separate the feature vectors with an optimal hyperplane in terms of margins, i.e., distances of given examples from a separating hyperplane. The generalization ability of SVMs has been analyzed, mainly in the framework of the PAC learning [6] where the VC dimension plays an important role [7]. Recently, studies on a more practical criterion, the average generalization error, have also been presented [8]–[12].

Another important topic regarding SVMs is how we should choose a kernel function, which has a well-defined feature space. Since any positive semidefinite function can be a kernel function, we can make a new kernel function K such as

$$K := \alpha_1 K_1 + \alpha_2 K_2 \quad (\alpha_1, \alpha_2 > 0) \tag{1}$$

$$K := K_1 K_2 \tag{2}$$

$$K := \exp[K_1] \tag{3}$$

from two arbitrary kernel functions K_1 and K_2 where $:=$ means definition. Hence we need to clarify which kernel is suitable for given data. Such a problem is called "learning kernels" and has been intensively studied. For example, [13] showed how to determine hyperparameters in a set of parametric kernel functions from the viewpoint of model selection in a Bayesian framework and [14] optimized the kernel function as a problem of transduction. However, both assumed a fixed set of parametric kernel functions and reduced the problem to parameter estimation. Therefore, the problem of learning kernels appears to remain open.

The purpose of this study is to contribute to this important problem by elucidating the effects of the properties of a kernel function on the SVM solutions. From this viewpoint, some asymptotic properties of SVM's with the Gaussian kernel have been reported in [15] when the two parameters of the kernel method, i.e., the steepness σ^2 of the kernel and the softness C of the margins, go to null or infinity. Although the results are important for learning kernels, some seem strange or unusual. For example, when the parameter C, which determines the significance of constraint violation, approaches null, all the examples are classified to the same category. This is caused by the formulation that positive and negative examples are separately treated, and they are unbalanced. Hence, as an alternative, we analyzed the so-called nu support vector machines (ν-SVM's), which are a variation of the SVM's proposed in [16] and which do not distinguish positive and negative examples from a geometrical viewpoint when homogeneous separating hyperplanes are assumed [12], [17].

In this paper, we do not restrict the kernel function to a parametric model but assume only its property in two extreme cases: One is that the diagonal elements are unity and the off-diagonal elements are almost null, that is, all feature vectors are almost orthogonal; The other is that the diagonal elements are unity and the off-diagonal elements are almost unity, that is, all feature vectors are almost the same. As a result, it is shown that the solution of the ν-SVM is nearly the center of gravity of the given examples in the former case while the solution is approximated to that of the ν-SVM with the linear kernel in the latter case. Although such extreme kernels are not employed in practice, analyses are helpful to understand the effects of a kernel function on the generalization performance.

The rest of the paper is organized as follows: Section II introduces the ν-SVM and its geometrical interpretation. We analyze asymptotic properties on the solution of ν-SVM in cases where the inner product of two distinct input vectors always takes very small or very large values; that is, null or unity in Section III. The results are applied to the cases discussed in [15] in Section IV, and are confirmed by computer simulations in Section V. Conclusions are given in Section VI.

II. The Nu SVMs

A. Formulation

SVMs are a kind of kernel methods; that is, they nonlinearly map an input vector $\boldsymbol{x}$ to a feature vector $\boldsymbol{f}(\boldsymbol{x})$ and separate the feature vector linearly in a high-dimensional feature

Manuscript received March 1, 2005; revised June 30, 2005. This work was supported by Grant-in-Aid for Scientific Research (14084210, 15700130) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

The author is with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: kazushi@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TNN.2005.860832

space. Among many hyperplanes in the feature space that correctly separate all given examples, an SVM chooses one that maximizes the margin defined as the minimum distance of examples from a separating hyperplane. We consider a homogeneous linear dichotomy in the feature space called a Perceptron, whose separating function is represented by $\boldsymbol{w}'\boldsymbol{f}(\boldsymbol{x})$ where $\boldsymbol{w} \in R^M$ is called the parameter vector and $'$ denotes the transpose. Note that an inhomogeneous linear dichotomy whose separating function is represented by $\boldsymbol{w}'\boldsymbol{f}(\boldsymbol{x}) + b$ is easily transformed to a homogeneous one $\tilde{\boldsymbol{w}}'\tilde{\boldsymbol{x}}$ by lifting-up; that is, by using augmented vectors $\tilde{\boldsymbol{w}} := (\boldsymbol{w}; b)$ and $\tilde{\boldsymbol{f}}(\boldsymbol{x}) := (\boldsymbol{f}(\boldsymbol{x}); 1)$ where $(\cdot;\cdot) := (\cdot', \cdot')'$.

According to the concept mentioned above, given N examples $(\boldsymbol{x}^{(n)}, y^{(n)}), n = 1, \ldots, N$, the ν-SVM proposed in [16] solves the following optimization problem:

$$\min_{\boldsymbol{w},\xi_n,\beta} \left[\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{n=1}^{N}\xi_n - \beta\right] \quad \text{s.t.} \boldsymbol{w}'\boldsymbol{f}^{(n)} \geq \beta - \xi_n, \quad \xi_n \geq 0 \quad (4)$$

where $\boldsymbol{f}^{(n)} := y^{(n)}\boldsymbol{f}(\boldsymbol{x}^{(n)})$ and C is a constant for soft margins [18]. Note that the original ν-SVM introduced in [16] is formulated using inhomogeneous separating hyperplanes as is in Appendix II and is of a slightly different form to (4). However, the equivalence between the original and (4) can easily be proven taking into account $1/\nu = CN$ as discussed in [12]. If the variable β is fixed to unity, (4) results in the original SVMs [1], [2].

The problem (4) is known to be equivalent to the following optimization problem called the dual problem:

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2 \quad \text{s.t. } \boldsymbol{w} = \sum_{n=1}^{N}\alpha_n\boldsymbol{f}^{(n)}, \quad 0 \leq \alpha_n \leq C, \textstyle\sum_{n=1}^{N}\alpha_n = 1 \quad (5)$$

where α_n are the Lagrange multipliers [12], [17].

One property of the kernel methods including the ν-SVM's is that the so-called kernel trick is applicable; that is, the inner product of a parameter vector $\boldsymbol{w}$ in (5) and a feature vector $\boldsymbol{f}(\boldsymbol{x})$ can be calculated without the explicit expression of feature vectors as follows:

$$\boldsymbol{w}'\boldsymbol{f}(\boldsymbol{x}) = \sum_{n=1}^{N}\alpha_n\boldsymbol{f}^{(n)\prime}\boldsymbol{f}(\boldsymbol{x}) \quad (6)$$

$$= \sum_{n=1}^{N}\alpha_n y^{(n)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right) \quad (7)$$

where $K(\cdot,\cdot)$ is a kernel function that determines the inner product in the feature space. In fact, Mercer's theorem shows that a nonlinear function $\boldsymbol{f}(\cdot)$ exists if and only if the kernel function is positive semidefinite. Therefore, choosing a kernel function means determining a nonlinear feature map and vice versa.

Another property is that the feature vector $\boldsymbol{f}(\boldsymbol{x}^{(n)})$ and the corresponding output $y^{(n)}$ of an example $\boldsymbol{x}^{(n)}$ always appear together in the form of $\boldsymbol{f}^{(n)} = y^{(n)}\boldsymbol{f}(\boldsymbol{x}^{(n)})$. This means that an example $(\boldsymbol{f}(\boldsymbol{x}^{(n)}), y^{(n)})$ is perfectly equivalent to $(-\boldsymbol{f}(\boldsymbol{x}^{(n)}), -y^{(n)})$ in (5) and hence we do not have to assume the imbalance of positive and negative examples. For this reason, we call $\boldsymbol{f}^{(n)}$ an example in the following. Note that if we regard all examples as chosen from one class, the formulation above is almost equivalent to the so-called one-class support vector machines [19] where the separating function is $\boldsymbol{w}'\boldsymbol{f}(\boldsymbol{x}) - \beta$ instead of $\boldsymbol{w}'\boldsymbol{f}(\boldsymbol{x})$.

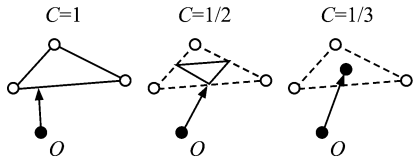

Fig. 1. Reduced convex hulls and the nearest points for $C = 1, C = 1/2$, and $C = 1/N$ when $N = 3$.

B. Reduced Convex Hull

One important advantage of the ν-SVMs in the SVM family is that (5) has a clear geometrical meaning. Minimizing the cost function $(1/2)||\boldsymbol{w}||^2$ is equivalent to finding the point $\hat{\boldsymbol{w}}$ nearest the origin that satisfies the constraints in (5). When $C = 1$, this restriction means that $\boldsymbol{w}$ belongs to the convex hull of $\{\boldsymbol{f}^{(n)}\}$, since the sum of nonnegative weights, α_n, is unity. For an arbitrary $C < 1$, the set in which $\hat{\boldsymbol{w}}$ can exist is reduced to the so-called reduced convex hull [20], [21] by the restriction $\alpha_n \leq C$; since one example can not contribute so much, $\hat{\boldsymbol{w}}$ should consist of more examples, e.g., two examples when $C = 1/2$. The reduced convex hull shrinks into the center of gravity of all the examples when $C = 1/N$ and vanishes when $C < 1/N$. So, the ν-SVM results in the problem of finding the point nearest the origin in the reduced convex hull (Fig. 1) [12], [17]. An example $\boldsymbol{f}^{(n)}$ that has a positive weight $\alpha_n > 0$ is called a support vector. The number of support vectors is related to the generalization ability in the framework of the PAC learning as Theorem 5.2 in [1]. We denote the set of indices of support vectors by V and its complement by $\overline{V}$.

C. Circumscribed Hypersphere

Suppose $||\boldsymbol{f}^{(n)}|| = 1$, that is, feature vectors are located on a hypersphere S, and consider hard-margins' case. Then, for any example $\boldsymbol{f}^{(n)}$, $\hat{\boldsymbol{w}}$ satisfies

$$\hat{\boldsymbol{w}}'\boldsymbol{f}^{(n)} \geq ||\hat{\boldsymbol{w}}||^2 \quad (8)$$

since $\hat{\boldsymbol{w}}$ is the nearest point and hence

$$\left\|\boldsymbol{f}^{(n)} - \hat{\boldsymbol{w}}\right\|^2 = \left\|\boldsymbol{f}^{(n)}\right\|^2 + ||\hat{\boldsymbol{w}}||^2 - 2\hat{\boldsymbol{w}}'\boldsymbol{f}^{(n)} \quad (9)$$

$$\leq \left\|\boldsymbol{f}^{(n)}\right\|^2 - ||\hat{\boldsymbol{w}}||^2 = \text{const.} \quad (10)$$

Since the equality holds when $\boldsymbol{f}^{(n)}$ is a support vector, the support vectors are equidistant from $\hat{\boldsymbol{w}}$; i.e., they lie on a circumscribed hypersphere centered at $\hat{\boldsymbol{w}}$, and the other examples are inside the hypersphere [22]. Since margin maximization corresponds to radial minimization, this hypersphere is the smallest that covers all examples (see Fig. 2), [17].

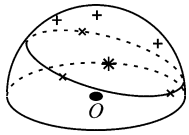

Fig. 2. The SVM solution (asterisk) corresponds to the center of the smallest ball including all examples in the feature space. Support vectors (crosses) are on the surface while the others (pluses) are inside on the unit hypersemisphere of the examples.

III. Asymptotic Properties of ν-SVM Solutions

We discuss the asymptotic properties of ν-SVMs under the assumption that feature vectors are normalized, $\|\boldsymbol{f}(\boldsymbol{x})\| = 1$.

In the first case, the kernel function takes a very small value, nearly null, for two distinct inputs. This means that any two input vectors are almost orthogonal in the feature space. The first subsection proves that the solution of the ν-SVM in this case is almost the gravity of center of the feature vectors.

In the second case, the kernel function takes a very large value, nearly unity due to the assumption of normalization, for two distinct input vectors. This means that any two input vectors are almost the same in the feature space. The second subsection shows that the solution of the ν-SVM in this case is approximated to that of the so-called linear kernel SVM with inhomogeneous separating hyperplanes.

A. Feature Vectors Are Almost Orthogonal

Suppose that given examples are not linearly separable. The soft margin technique is one method to treat such a problem [18]; however, it sometimes reduces the generalization ability [12]. Another is to employ a kernel function that makes examples linearly separable in the feature space. For example, the Gaussian kernel has an infinite-dimensional feature space and all the feature vectors of given examples are linearly independent. This means that any set of examples becomes linearly separable in the feature space.

The condition that feature vectors are almost orthogonal is the ultimate in the latter case. Since feature vectors are little correlated to each other, it is expected that the ν-SVM has a low generalization ability. In fact, the following theorem can be proven as below in this case:

Theorem 1: If the kernel function takes a small value for any pair of examples, that is

$$\left|K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}\right)\right| < \delta := \frac{1}{4N} \tag{11}$$

for any $n \neq m$, then any feature vector is a support vector of the ν-SVM with hard margins.

Proof: Given N examples $\boldsymbol{f}^{(n)}, n = 1, \ldots, N$, the solution of the ν-SVM is the center $\boldsymbol{c}_N^*$ of the minimum ball that includes all examples and is written as

$$\hat{\boldsymbol{w}} = \sum_{n \in V}^{N} \hat{\alpha}_n \boldsymbol{f}^{(n)} \tag{12}$$

where $\hat{\alpha}_n > 0$ and $\sum \hat{\alpha}_n = 1$. From the discussion in Subsection II.C, $\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(n)}\|$ for $n \in V$ is constant and satisfies

$$\left\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(n)}\right\| > \max_{m \in \bar{V}} \left\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(m)}\right\|. \tag{13}$$

Therefore, if we show

$$\left\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(n)}\right\| < \left\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(m)}\right\| \tag{14}$$

for $m \in \bar{V}$, this means $\bar{V} = \emptyset$ and the proof is completed.

Let $\boldsymbol{c}$ and n^* be the center of gravity of all support vectors and the index of the farthest support vector from $\boldsymbol{c}$, respectively; that is

$$\boldsymbol{c} := \frac{1}{M} \sum_{n \in V} \boldsymbol{f}^{(n)}, \tag{15}$$

$$n^* := \arg\max_{n \in V} \left\|\boldsymbol{c} - \boldsymbol{f}^{(n)}\right\|^2 \tag{16}$$

where M is the number of support vectors, $M := |V|$. Then, since the ball with center $\boldsymbol{c}$ and radius $\|\boldsymbol{c} - \boldsymbol{f}^{(n^*)}\|$ includes all the support vectors, the left-hand side of (14) satisfies

$$\begin{aligned}
&\left\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(n)}\right\|^2 \\
&\leq \left\|\boldsymbol{c} - \boldsymbol{f}^{(n^*)}\right\|^2 && (17)\\
&= \left\|\left(1 - \frac{1}{M}\right)\boldsymbol{f}^{(n^*)} - \sum_{n \in V - \{n^*\}} \frac{1}{M}\boldsymbol{f}^{(n)}\right\|^2 && (18)\\
&\leq \left(1 - \frac{1}{M}\right)^2 + \frac{M-1}{M^2} + \delta\frac{M(M-1)}{M^2} \\
&\quad + 2\delta\frac{M-1}{M}\left(1 - \frac{1}{M}\right) && (19)\\
&= 1 - \frac{1}{M} + \frac{\delta(M-1)(3M-2)}{M^2} && (20)\\
&\leq 1 - \frac{1}{4M} && (21)
\end{aligned}$$

while the right-hand side satisfies

$$\begin{aligned}
\left\|\hat{\boldsymbol{w}} - \boldsymbol{f}^{(m)}\right\|^2 &= 1 + \sum_{n \in V} \hat{\alpha}_n^2 - 2\boldsymbol{f}^{(m)} \\
&\quad \cdot \sum_{n \in V} \hat{\alpha}_n \boldsymbol{f}^{(n)} + \sum_{n,l \in V, n \neq l} \hat{\alpha}_n \hat{\alpha}_l \boldsymbol{f}^{(n)} \cdot \boldsymbol{f}^{(l)} && (22)\\
&\geq 1 + \sum_{n \in V} \hat{\alpha}_n^2 - 2\delta \\
&\quad - M(M-1)\left(\frac{1}{M} + 2\delta\right)^2 \delta && (23)\\
&\geq 1 - \frac{1}{16M}. && (24)
\end{aligned}$$

Here, we have used $M \leq N$ and

$$\sum_{n \in V} \hat{\alpha}_n^2 \geq \frac{1}{M} \tag{25}$$

$$\hat{\alpha}_n \leq \frac{1}{M} + 2\delta \tag{26}$$

which are proven in Appendix I. Hence (14) holds true for any M from (21) and (24). ■

Corollary 2: Theorem 1 holds for the ν-SVM with soft margins if $C \geq (3/2N)$.

This is proven from the facts that $\boldsymbol{w}$ approaches $\boldsymbol{c}$ and that $\hat{\boldsymbol{\alpha}}$ satisfies (26).

B. Feature Vectors Are Almost the Same

Another ultimate in the opposite direction of the previous subsection is the case where the kernel function takes a very large value, nearly unity for any distinct input vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$,

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) \approx 1. \tag{27}$$

In this case, the set of examples $\boldsymbol{f}^{(n)} = y^{(n)}\boldsymbol{f}(\boldsymbol{x}^{(n)})$ is divided into two groups; each example in one group has inner products of nearly one with the examples in the same group and those of nearly minus one with the others (see Fig. 3). That is, the property of the ν-SVM that it does not distinguish positive and negative examples is lost here.

Since analyzing general cases is so difficult, we assume in the following that the kernel function has the form:

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) := \Psi\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{\sigma^2}\right) \tag{28}$$

where $\Psi(x)$ is a monotonically decreasing differentiable function with $\Psi(0) = 1, \Psi(\infty) = 0$ and $\Psi'(0) = -D_0 < 0$, and σ takes a large value. One example is the Gaussian kernel where $\Psi(x) = \exp(-x)$ and another example is the polynomial kernel with normalized inputs. Using the Taylor expansion, such a kernel function is written as

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) = K(0) - \frac{D_0}{\sigma^2}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 + \frac{D(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sigma^4}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^4 \tag{29}$$

where

$$D(\boldsymbol{x}_1, \boldsymbol{x}_2) := \frac{1}{2}\Psi''\left(\eta\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{\sigma^2}\right) \tag{30}$$

and $\eta \in [0, 1]$ depending on $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2$. We set $K(0) = 1$ and $R = \max\|\boldsymbol{x}^{(n)}\|$ without loss of generality, and assume that

$$\epsilon(\boldsymbol{x}_1, \boldsymbol{x}_2) := \frac{D(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sigma^4}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^4 \tag{31}$$

has an upper-bound ϵ in magnitude, i.e., $|\epsilon(\boldsymbol{x}_1, \boldsymbol{x}_2)| \leq \epsilon$ for any input vectors $\boldsymbol{x}_1, \boldsymbol{x}_2$. In this case, the ν-SVM is approximated to the inhomogeneous ν-SVM (the original ν-SVM) with the linear kernel $K(\boldsymbol{x}_1, \boldsymbol{x}_2) := \boldsymbol{x}_1'\boldsymbol{x}_2$, that is, $\boldsymbol{f}(\boldsymbol{x}) := \boldsymbol{x}$.

From the assumption (28), the cost function $E(\boldsymbol{\alpha}) = \|\boldsymbol{w}\|^2$ of the ν-SVM is expressed as

$$E(\alpha) := \sum_{n,m}\alpha_n\alpha_m y^{(n)}y^{(m)}K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}) \tag{32}$$

$$= \left(\sum_n \alpha_n y^{(n)}\right)^2 - \sum_{n,m}\alpha_n\alpha_m y^{(n)}y^{(m)}\frac{D_0}{\sigma^2}\|\boldsymbol{x}^{(n)} - \boldsymbol{x}^{(m)}\|^2 + \sum_{n,m}\alpha_n\alpha_m y^{(n)}y^{(m)}\frac{\epsilon_{nm}}{\sigma^4} \tag{33}$$

$$= \left(\sum_n \alpha_n y^{(n)}\right)^2 + 2\sum_{n,m}\alpha_n\alpha_m y^{(n)}y^{(m)}\frac{D_0}{\sigma^2}\boldsymbol{x}^{(n)\prime}\boldsymbol{x}^{(m)} - 2\sum_m \alpha_m y^{(m)}\sum_n \alpha_n y^{(n)}\frac{D_0}{\sigma^2}\left\|\boldsymbol{x}^{(n)}\right\|^2 + \sum_{n,m}\alpha_n\alpha_m y^{(n)}y^{(m)}\frac{\epsilon_{nm}}{\sigma^4} \tag{34}$$

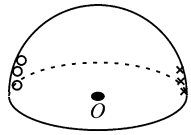

Fig. 3. Examples are divided into two groups in the feature space. One consists of positive examples (o's) and the other of negative examples (x's).

where $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_N)$ is called a coefficient vector and ϵ_{nm} stands for $\epsilon(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)})$. The solution of the ν-SVM is the minimizer of $E(\boldsymbol{\alpha})$, denoted by $\hat{\boldsymbol{\alpha}} := (\hat{\alpha}_1, \ldots, \hat{\alpha}_N)$. Note that the last term of (33) is $\mathrm{O}(\sigma^{-4})$ while the second term of (33) is $\mathrm{O}(\sigma^{-2})$. This means that the last term of (33) little affects the value of the cost function when σ takes a large value. One property of the ν-SVM is that any example $(\boldsymbol{x}^{(l)}, y^{(l)})$ s.t. $\hat{\alpha}_l \neq 0$ satisfies

$$y^{(l)}\sum_n \hat{\alpha}_n y^{(n)}K\left(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(n)}\right) = E(\hat{\boldsymbol{\alpha}}) = \sum_{n,m}\hat{\alpha}_n\hat{\alpha}_m y^{(n)}y^{(m)}K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}) \tag{35}$$

as mentioned in Subsection II.C.

We introduce another coefficient vector denoted by $\boldsymbol{\alpha}^{\mathrm{P}} = (\alpha_1^{\mathrm{P}}, \ldots, \alpha_N^{\mathrm{P}})$ that satisfies $\sum_n \alpha_n^{\mathrm{P}} y^{(n)} = 0$. Denote the set of indices for positive examples by V_+ and that for negative examples by V_-

$$V_+ := \{n | y^{(n)} = +1\}, \quad V_- := \{n | y^{(n)} = -1\} \tag{36}$$

and define $\boldsymbol{\alpha}^{\mathrm{P}} = (\alpha_1^{\mathrm{P}}, \ldots, \alpha_N^{\mathrm{P}})$ as

$$\alpha_n^{\mathrm{P}} := \frac{1}{1 + y^{(n)}\Delta}\hat{\alpha}_n \tag{37}$$

using the solution $\hat{\boldsymbol{\alpha}}$ of the ν-SVM and $\Delta := \sum_n \hat{\alpha}_n y^{(n)}$. We can easily confirm that α^{P} satisfies

$$\sum_{n\in V_+}\alpha_n^{\mathrm{P}} = \sum_{n\in V_-}\alpha_n^{\mathrm{P}} = \frac{1}{2} \tag{38}$$

and hence

$$\sum_n \alpha_n^{\mathrm{P}} y^{(n)} = 0 \tag{39}$$

$$\sum_n \alpha_n^{\mathrm{P}} = 1 \tag{40}$$

since

$$\sum_{n\in V_+}\hat{\alpha}_n = \frac{1 + \Delta}{2} \tag{41}$$

$$\sum_{n\in V_-}\hat{\alpha}_n = \frac{1 - \Delta}{2}. \tag{42}$$

In order to see how well $\boldsymbol{\alpha}^{\mathrm{P}}$ can approximate to $\boldsymbol{\alpha}$ in separating input vectors as σ tends to infinity, we evaluate the difference of their outputs for an arbitrary input vector $\boldsymbol{x}$. Using the fact that Δ is $\mathrm{O}(\sigma^{-2})$ in magnitude, as shown in Appendix III

$$\sum_n \hat{\alpha}_n y^{(n)}K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right) - \sum_n \alpha_n^{\mathrm{P}} y^{(n)}K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right) \tag{43}$$

$$= \Delta\sum_n \frac{\hat{\alpha}_n}{1 + y^{(n)}\Delta}K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right) \tag{44}$$

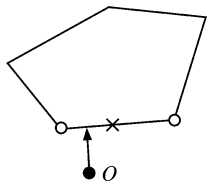

Fig. 4. Support vector (x) is expressed as a linear combination of other support vectors (o's) when they are in the same hyperplane. Here, a polygon represents the convex hull of examples.

$$= \Delta + \Delta\frac{2D_0}{\sigma^2}\sum_n \frac{\hat{\alpha}_n}{1+y^{(n)}\Delta}\boldsymbol{x}^{(n)\prime}\boldsymbol{x} - \Delta\frac{D_0}{\sigma^2}\sum_n \frac{\hat{\alpha}_n}{1+y^{(n)}\Delta}\left\|\boldsymbol{x}^{(n)}\right\|^2 - \Delta\frac{D_0}{\sigma^2}\|\boldsymbol{x}\|^2 + \Delta\mathrm{O}(\sigma^{-4}) \quad (45)$$

$$= \Delta + \mathrm{O}(\sigma^{-4}). \quad (46)$$

Hence, the difference is always Δ with the precision of $\mathrm{O}(\sigma^{-2})$ and we can substitute $\boldsymbol{\alpha}^{\mathrm{P}}$ for $\hat{\boldsymbol{\alpha}}$ by adding Δ.

Next, we see the difference of the outputs of $\boldsymbol{\alpha}^{\mathrm{P}}$ and the solution of the inhomogeneous ν-SVM with the linear kernel. The latter is the minimizer of

$$E^{\mathrm{I}}(\boldsymbol{\alpha}) := \frac{2D_0}{\sigma^2}\sum_{n,m}\alpha_n\alpha_m y^{(n)}y^{(m)}\boldsymbol{x}^{(n)\prime}\boldsymbol{x}^{(m)} \quad \text{s.t. } 0 \le \alpha_n \le C,$$

$$\sum_{n=1}^{N}\alpha_n = 1, \quad \sum_n \alpha_n y^{(n)} = 0 \quad (47)$$

and is denoted by $\boldsymbol{\alpha}^{\mathrm{I}} := (\alpha_1^{\mathrm{I}}, \ldots, \alpha_N^{\mathrm{I}})$.

In the following, we assume the uniqueness of $\boldsymbol{\alpha}^{\mathrm{I}}$. This assumption is made to avoid peculiar cases such that a support vector is expressed as a linear combination of other support vectors (see Fig. 4). If we remove support vectors expressed as a positively weighted sum of other examples, this is unnecessary.

Since $|\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}|$ has an upper-bound of $\mathrm{O}(\sigma^{-1})$ as shown in Appendix IV, the difference for an arbitrary input vector $\boldsymbol{x}$ is at most $\mathrm{O}(\sigma^{-3})$ because

$$\left|\sum_n \alpha_n^{\mathrm{P}} y^{(n)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right) - \sum_n \alpha_n^{\mathrm{I}} y^{(n)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right)\right| \quad (48)$$

$$\le \sum_n \left|\left(\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}\right)\frac{D_0}{\sigma^2} y^{(n)}\boldsymbol{x}^{(n)\prime}\boldsymbol{x}\right| + \sum_n \left|\left(\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}\right)\frac{4\epsilon}{\sigma^4}\right| \quad (49)$$

$$\le \frac{C_2 D_0 N R^2}{\sigma^3} + \frac{4C_2\epsilon N}{\sigma^5} \quad (50)$$

$$= \mathrm{O}(\sigma^{-3}). \quad (51)$$

Hence, the difference can be neglected compared to the magnitude of the output for a support vector in (35), which is $\mathrm{O}(\sigma^{-2})$.

Combining both differences, it has been shown that the ν-SVM for a large σ is approximated to the inhomogeneous ν-SVM with the linear kernel by adding Δ. Note that in this case there exists no ν-SVM solution unless given examples are linearly separable in the input space.

IV. Example: Gaussian Kernel

We apply the results in the previous section to the ν-SVM with the Gaussian kernel and consider five asymptotic cases discussed in [15]. Note that the parameter C in the ν-SVM is meaningful only when $1/N \le C \le 1$; There exists no solution when $C < 1/N$ and the solution for $C \ge 1$ is the same as that for $C = 1$, whereas the conventional SVM has a unique solution for any example set and any positive parameters, C and σ^2. Therefore, we consider the asymptotic $C \to 1/N$ instead of $C \to 0$.

Case 1: σ^2 Fixed and $C \to 1/N$: The solution by the original SVM underfits the given examples and classifies any of them to the majority class as $C \to 0$ since the slack variables ξ_n are ignored in the cost function [15]. The solution by the ν-SVM, on the other hand, converges to the center of gravity of the given examples as $C \to 1/N$, since the reduced convex hull reduces to the point, as discussed in Subsection II-B; that is

$$\hat{\boldsymbol{\alpha}} \to \frac{1}{N}\mathbf{1} := \frac{1}{N}(1, \ldots, 1). \quad (52)$$

This is a kind of underfitting since the separating hyperplane is chosen without taking into account whether the given examples are correctly classified or not.

Case 2: σ^2 Fixed and $C \to \infty$: The solution by the original SVM classifies any of them as correctly as possible and it approaches the solution of the hard margin problem as $C \to \infty$ [15]. The solution by the ν-SVM is the same as that with hard margins ($C = 1$) in the same way, since the reduced convex hull with $C \ge 1$ coincides to that with $C = 1$.

Case 3: C Fixed and $\sigma^2 \to 0$: The solution by the original SVM overfits or underfits given examples depending on C as $\sigma^2 \to 0$ [15]. In the ν-SVM, however, this case satisfies the condition of Theorem 1 since $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \to \delta_{ij}$ as $\sigma^2 \to 0$, and hence the solution by the ν-SVM converges to the center of gravity as discussed in Subsection III-A.

Note that we cannot simply say whether this is overfitting or underfitting because this is a kind of overfitting since the separating hyperplane is chosen so that all given examples are separated correctly, and this is a kind of underfitting since any example contributes only $\alpha_n = 1/N$ at the same time.

Case 4: C Fixed and $\sigma^2 \to \infty$: Since $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \to 1$ as $\sigma^2 \to \infty$, this case satisfies the condition discussed in Subsection III-B and hence the solution of the ν-SVM approaches that of the inhomogeneous ν-SVM with the linear kernel.

This result is similar to that in [15]. One difference is that $\Delta = \sum_n \hat{\alpha}_n y^{(n)}$ is always null in the original SVM while it approaches null in the ν-SVM. Due to these properties, either of them can be approximated with an SVM with the linear kernel. Another difference is that the approximation by the linear kernel does not relate to the effect of C in the ν-SVM while C works as if C/σ^2 in the original SVM. This is because C appears only as the constraints in the reduced convex hull in the ν-SVM while $\sum_n \alpha_n$ varies according to C in the original SVM.

Case 5: $C \to \infty$ and $\sigma^2 \to \infty$ With a Fixed Ratio: Since the effect of C does not change when C becomes greater than 1, this is equivalent to the previous case in the ν-SVM. The original SVM converges to the solution of the inhomogeneous SVM with the linear kernel with C/σ^2 instead of C.

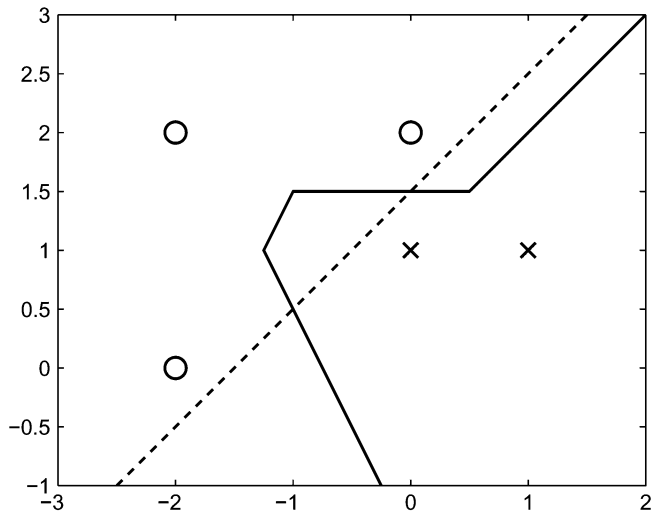

Fig. 5. Three positive examples (o's), $\boldsymbol{x}^{(1)} = (0, 2), \boldsymbol{x}^{(2)} = (-2, 0), \boldsymbol{x}^{(3)} = (-2, 2)$, and two negative examples (x's), $\boldsymbol{x}^{(4)} = (0, 1), \boldsymbol{x}^{(5)} = (1, 1)$. The solid and dashed lines express the decision boundaries for $\sigma \to 0$ and for $\sigma \to \infty$, respectively.

TABLE I
COEFFICIENT VECTORS OF ν-SVMS WITH THE GAUSSIAN KERNEL

σ^2	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$
.1	.2000	.2000	.2000	.2000	.2000
.3	.2072	.1998	.1998	.2003	.1929
1	.2814	.1779	.1728	.2375	.1303
2	.3505	.1479	.1005	.3306	.0704
3	.3854	.1454	.0438	.4019	.0236
5	.4023	.1487	0	.4489	0
10	.3925	.1383	0	.4692	0
100	.3771	.1265	0	.4963	0
1000	.3752	.1252	0	.4996	0
Center of Gravity	.2	.2	.2	.2	.2
Linear Kernel	.375	.125	0	.5	0

V. COMPUTER SIMULATIONS

To confirm the validity of the theoretical analysis given above, some computer simulations were carried out. In each experiment, the ν-SVM with two-dimensional (2-D) input space is given three positive examples, $\boldsymbol{x}^{(1)} = (0, 2), \boldsymbol{x}^{(2)} = (-2, 0), \boldsymbol{x}^{(3)} = (-2, 2)$, and two negative examples, $\boldsymbol{x}^{(4)} = (0, 1), \boldsymbol{x}^{(5)} = (1, 1)$, as shown in Fig. 5. The kernel function is Gaussian, that is

$$K(\boldsymbol{x}_1, \boldsymbol{x}_2) := \exp\left(-\frac{||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2}{\sigma^2}\right) \tag{53}$$

and the parameter σ^2 varies from .1 to 1000. The coefficient vector $\hat{\boldsymbol{\alpha}}$ of the ν-SVM solution for each σ^2 is given in Table I. When σ^2 is small, the vector corresponds to $\mathbf{1}/N$ and it approaches the coefficient vector of the inhomogeneous ν-SVM with the linear kernel as σ^2 increases, as proven in Section III.

VI. CONCLUSION

We discussed the properties of the solution of the ν-SVM in extreme cases. When all feature vectors are almost orthogonal, the solution of the ν-SVM is nearly the center of gravity of the examples where the coefficient vector is $\mathbf{1}/N$. Contrarily, when feature vectors are almost the same, the solution is approximated to that of the inhomogeneous ν-SVM with the linear kernel. These results were confirmed by computer simulations.

These results imply that the ν-SVM has a low generalization ability in both cases as below. Although the leave-one-out (LOO) error employed in [15] is a good approximate of the average generalization error, we cannot calculate it in the formulation of this paper. Instead, we consider in [1, Th. 5.2] here, although this holds only for the original SVM and not necessarily for the ν-SVM with homogeneous hyperplanes. The average generalization error has three upper-bounds

$$\frac{M}{N}, \frac{R^2}{N||\hat{\boldsymbol{w}}||^2}, \frac{L}{N} \tag{54}$$

where L is the dimension of the feature space. When all feature vectors are almost orthogonal, that is, the kernel function satisfies $|K(\cdot, \cdot)| \leq 1/(4N)$, then

$$\frac{M}{N} = 1 \tag{55}$$

$$\frac{R^2}{N||\hat{\boldsymbol{w}}||^2} \geq \frac{4R^2}{7} \tag{56}$$

$$\frac{L}{N} \geq 1 \tag{57}$$

since

$$||\hat{\boldsymbol{w}}||^2 = \sum_{n,m} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}\right) \tag{58}$$

$$\leq \sum_{n,m} \hat{\alpha}_n \hat{\alpha}_m \delta + (1 - \delta) \sum_n \alpha_n^2 \leq \frac{1}{4N} + \frac{3}{2N} \tag{59}$$

where we apply (64). Therefore, its generalization ability is very low. Contrarily, when feature vectors are almost the same, the generalization ability is the same as that of the inhomogeneous ν-SVM with the linear kernel, which means that we fail to choose a kernel function in a sense.

This paper has analyzed under some restrictive assumptions what happens if we employ extreme kernel functions. Although the results are interesting from the theoretical viewpoint, they little contribute practitioners since we could expect such kernel functions would be useless. Hence, providing a more practical guideline for kernel selection remains for future work.

APPENDIX I
PROOF OF $\alpha_n \leq 1/M + 2\delta$

(25) is easily shown from the nonnegativeness of α_n. (26) is proven as below.

Since $||\hat{\boldsymbol{w}} - \boldsymbol{f}^{(m)}||$ is constant for any $m \in V$, so is $\hat{\boldsymbol{w}} \cdot \boldsymbol{f}^{(m)}$, which we denote by a. Then we have

$$\hat{\alpha}_m = -\boldsymbol{f}^{(m)\prime} \sum_{n \in V, n \neq m} \hat{\alpha}_n \boldsymbol{f}^{(n)} + a. \tag{60}$$

Since the inner product of distinct examples is less than δ in magnitude and $\sum_{n \in V} \hat{\alpha}_n = 1$,

$$a - \delta \leq \hat{\alpha}_m \leq a + \delta \tag{61}$$
$$\hat{\alpha}_m - \delta \leq a \leq \hat{\alpha}_m + \delta \tag{62}$$

hold true. Summing up (62) for all $m \in V$, we get

$$\frac{1}{M} - \delta \leq a \leq \frac{1}{M} + \delta \tag{63}$$
$$\frac{1}{M} - 2\delta \leq \hat{\alpha}_i \leq \frac{1}{M} + 2\delta \tag{64}$$

using (61) again. This means that when δ approaches null, $\hat{\alpha}_n$ converges to $1/M$ and hence $\hat{\boldsymbol{w}}$ to the center of gravity $\boldsymbol{c}$.

Appendix II
Formulation of the Inhomogeneous ν-SVM

The original ν-SVM's proposed in [16] employs inhomogeneous hyperplanes $\boldsymbol{w}'\boldsymbol{f}(\boldsymbol{x}) + b = 0$; that is, hyperplanes do not necessarily include the origin. Hence, we call it the inhomogeneous ν-SVM and formulate it as

$$\min_{\boldsymbol{w}, b, \xi_n, \beta} \left[\frac{1}{2}||\boldsymbol{w}||^2 + C \sum_{n=1}^{N} \xi_n - \beta \right]$$
$$\text{s.t. } y^{(n)} \left(\boldsymbol{w}'\boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right) + b \right) \geq \beta - \xi_n, \quad \xi_n \geq 0. \tag{65}$$

Its dual problem is easily derived to

$$\min_{\boldsymbol{w}} \frac{1}{2}||\boldsymbol{w}||^2$$
$$\text{s.t. } \boldsymbol{w} = \textstyle\sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{f}\left(\boldsymbol{x}^{(n)}\right), \quad 0 \leq \alpha_n \leq C,$$
$$\sum_{n=1}^{N} \alpha_n = 1, \quad \sum_{n=1}^{N} y^{(n)} \alpha_n = 0 \tag{66}$$

where α_n are the Lagrange multipliers. Note that $y^{(n)}$ appears alone in the last equation of (66) unlike (5).

If we define $\tilde{\boldsymbol{w}} = (\boldsymbol{w}; b) \in \tilde{F}$ and $\tilde{\boldsymbol{f}} = (\boldsymbol{f}; 1) \in \tilde{F}$ where $\tilde{F}$ is the augmented feature space $\{(\boldsymbol{f}, z)\} = F \times R$, the separating hyperplane is expressed as a simple inner product $\tilde{\boldsymbol{w}}'\tilde{\boldsymbol{f}} = 0$; that is, the hyperplane is homogeneous. This operation is called lifting-up (Fig. 6). Since (66) is the same as the dual problem of the homogeneous ν-SVM in (5) except for the last equation in (66), the solution $\hat{\boldsymbol{w}}$ of the inhomogeneous SVM is expressed as the point $(\hat{\boldsymbol{w}}; 0)$ nearest the origin in the intersection of the hyperplane $z = 0$ and the reduced convex hull of the lifted-up vectors $y^{(n)}(\boldsymbol{f}(\boldsymbol{x}^{(n)}); 1)$.

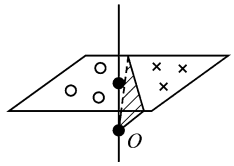

Fig. 6. Geometrical meaning of lifting-up. An inhomogeneous hyperplane (dashed line in a plane) in a space is expressed as a homogeneous hyperplane (hatched triangle) in an augmented space.

Appendix III
An Upper-Bound of Δ

Summing up (35) for all l with weight α_l^{P}, we get

$$\sum_{n,l} \hat{\alpha}_n \alpha_l^{\mathrm{P}} y^{(n)} y^{(l)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(l)}\right) = \sum_{n,m} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}\right) \tag{67}$$

that is

$$\sum_{n,m} \frac{y^{(n)}\Delta}{1 + y^{(n)}\Delta} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} K(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}) = 0. \tag{68}$$

Using (29), the above equation is rewritten as

$$\begin{aligned} 0 = &\sum_{n,m} \frac{y^{(n)}}{1 + y^{(n)}\Delta} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} \\ &+ 2\frac{D_0}{\sigma^2} \sum_{n,m} \frac{y^{(n)}}{1 + y^{(n)}\Delta} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} \\ &- \frac{D_0}{\sigma^2} \sum_{n,m} \frac{y^{(n)}}{1 + y^{(n)}\Delta} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} ||\boldsymbol{x}^{(n)}||^2 \\ &- \frac{D_0}{\sigma^2} \sum_{n,m} \frac{y^{(n)}}{1 + y^{(n)}\Delta} \hat{\alpha}_n \hat{\alpha}_m y^{(n)} y^{(m)} ||\boldsymbol{x}^{(m)}||^2 \\ &+ \sum_{n,m} \frac{y^{(n)}}{1 + y^{(n)}\Delta} \hat{\alpha}_n \hat{\alpha}_m \frac{\epsilon_{nm}}{\sigma^4}. \end{aligned} \tag{69}$$

Let the first term of the right-hand side of the equation be denoted by T_1, the second by T_2, the third by T_3, the fourth by T_4 and the fifth by T_5. Then, the following equations and inequalities hold true:

$$T_1 = \Delta \sum_n \frac{\hat{\alpha}_n}{1 + y^{(n)}\Delta} = \Delta \tag{70}$$
$$|T_2| \leq 2\frac{D_0}{\sigma^2} \sum_m \hat{\alpha}_m \sum_n \frac{\hat{\alpha}_n}{1 + y^{(n)}\Delta} R^2 = 2\frac{D_0}{\sigma^2} R^2 \tag{71}$$
$$|T_3| \leq \frac{D_0}{\sigma^2} \left| \sum_m \hat{\alpha}_m y^{(m)} \right| \sum_n \frac{\hat{\alpha}_n}{1 + y^{(n)}\Delta} R^2 = |\Delta| \frac{D_0}{\sigma^2} R^2 \tag{72}$$
$$|T_4| \leq \frac{D_0}{\sigma^2} \sum_n \frac{\hat{\alpha}_n}{1 + y^{(n)}\Delta} \sum_m \hat{\alpha}_m R^2 = \frac{D_0}{\sigma^2} R^2 \tag{73}$$

$$C_1 := \inf_{\boldsymbol{\alpha}^{\mathrm{P}}} \frac{\sum_{n,m}(\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}})(\alpha_m^{\mathrm{P}} - \alpha_m^{\mathrm{I}}) y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)}}{||\boldsymbol{\alpha}^{\mathrm{P}} - \boldsymbol{\alpha}^{\mathrm{I}}||^2}. \tag{83}$$

$$|T_5| \leq \sum_n \frac{\hat{\alpha}_n}{1 + y^{(n)}\Delta} \sum_m \hat{\alpha}_m \frac{\epsilon}{\sigma^4} = \frac{\epsilon}{\sigma^4} \tag{74}$$

which lead (68) to

$$|\Delta| \left(1 - \frac{D_0}{\sigma^2} R^2\right) \leq 3\frac{D_0}{\sigma^2} R^2 + \frac{\epsilon}{\sigma^4}. \tag{75}$$

Therefore, the magnitude of Δ is at most $\mathrm{O}(\sigma^{-2})$.

Appendix IV
An Upper-Bound of $|\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}|$

Using (29), an upper-bound of $E(\boldsymbol{\alpha}^{\mathrm{P}} - \boldsymbol{\alpha}^{\mathrm{I}})$ is given by

$$\sum_{n,m} \left(\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}\right) \left(\alpha_m^{\mathrm{P}} - \alpha_m^{\mathrm{I}}\right) y^{(n)} y^{(m)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}\right) \tag{76}$$

$$\leq \frac{2D_0}{\sigma^2} \sum_{n,m} (\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}})(\alpha_m^{\mathrm{P}} - \alpha_m^{\mathrm{I}}) y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} + \frac{4\epsilon}{\sigma^4} \tag{77}$$

$$\leq \frac{2D_0}{\sigma^2} \sum_{n,m} \alpha_n^{\mathrm{P}} \alpha_m^{\mathrm{P}} y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} - \frac{2D_0}{\sigma^2} \sum_{n,m} \alpha_n^{\mathrm{I}} \alpha_m^{\mathrm{I}} y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} + \frac{4\epsilon}{\sigma^4} \tag{78}$$

$$= E^{\mathrm{I}}(\boldsymbol{\alpha}^{\mathrm{P}}) - E^{\mathrm{I}}(\boldsymbol{\alpha}^{\mathrm{I}}) + \frac{4\epsilon}{\sigma^4} \leq \frac{6\epsilon}{\sigma^4}. \tag{79}$$

On the contrary, $E(\boldsymbol{\alpha}^{\mathrm{P}} - \boldsymbol{\alpha}^{\mathrm{I}})$ has a lower-bound such that

$$\sum_{n,m} (\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}) \left(\alpha_m^{\mathrm{P}} - \alpha_m^{\mathrm{I}}\right) y^{(n)} y^{(m)} K\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}\right) \tag{80}$$

$$\geq \frac{2D_0}{\sigma^2} \sum_{n,m} \left(\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}\right) \left(\alpha_m^{\mathrm{P}} - \alpha_m^{\mathrm{I}}\right) y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} - \frac{4\epsilon}{\sigma^4} \tag{81}$$

$$= C_1 \frac{2D_0}{\sigma^2} ||\boldsymbol{\alpha}^{\mathrm{P}} - \boldsymbol{\alpha}^{\mathrm{I}}||^2 - \frac{4\epsilon}{\sigma^4} \tag{82}$$

where (83) is shown at the top of the page. Note that C_1 is positive since

$$\sum_{n,m} (\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}})(\alpha_m^{\mathrm{P}} - \alpha_m^{\mathrm{I}}) y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} \geq \sum_{n,m} \alpha_n^{\mathrm{P}} \alpha_m^{\mathrm{P}} y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} - \sum_{n,m} \alpha_n^{\mathrm{I}} \alpha_m^{\mathrm{I}} y^{(n)} y^{(m)} \boldsymbol{x}^{(n)\prime} \boldsymbol{x}^{(m)} > 0 \tag{84}$$

from the uniqueness of the minimizer of $E^{\mathrm{I}}(\boldsymbol{\alpha})$. Hence, there exists a positive C_2 that satisfies

$$\max_n \left|\alpha_n^{\mathrm{P}} - \alpha_n^{\mathrm{I}}\right| \leq ||\boldsymbol{\alpha}^{\mathrm{P}} - \boldsymbol{\alpha}^{\mathrm{I}}|| \leq \frac{C_2}{\sigma}. \tag{85}$$

Acknowledgment

The authors would like to thank T. Komoto for his assistance with computer simulations.

References

[1] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[2] ——, *Statistical Learning Theory*. New York: John Wiley and Sons, 1998.
[3] B. Schölkopf, C. Burges, and A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*. Cambridge, Cambridgeshire, UK: Cambridge Univ. Press, 1998.
[4] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[5] A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, Eds., *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 2000.
[6] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, pp. 1134–1142, 1984.
[7] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Prob. Applicat.*, vol. 16, pp. 264–280, 1971.
[8] R. Dietrich, M. Opper, and H. Sompolinsky, "Statistical mechanics of support vector networks," *Phys. Rev. Lett.*, vol. 82, no. 14, pp. 2975–2978, 1999.
[9] M. Opper and R. Urbanczik, "Universal learning curves of support vector machines," *Physical Review Letters*, vol. 86, no. 19, pp. 4410–4413, 2001.
[10] K. Ikeda, "Geometry and learning curves of kernel methods with polynomial kernels," *Syst. Comput. Japan*, vol. 35, no. 7, pp. 41–48, 2004.
[11] ——, "An asymptotic statistical theory of polynomial kernel methods," *Neural Comput.*, vol. 16, no. 8, pp. 1705–1719, 2004.
[12] K. Ikeda and T. Aoishi, "An asymptotic statistical analysis of support vector machines with soft margins," *Neural Netw.*, vol. 18, no. 3, pp. 251–259, 2005.
[13] C. Gold and P. Sollich, "Model selection for support vector machine classification," *Neurocomputing*, vol. 55, pp. 221–249, 2003.
[14] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semi-definite programming," *J. Machine Learning Res.*, vol. 5, pp. 27–72, 2004.
[15] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, 2003.
[16] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Comput.*, vol. 12, no. 5, pp. 1207–1245, 2000.
[17] K. Ikeda and N. Murata, "Geometrical properties of nu support vector machines with different norms," *Neural Comput.*, vol. 17, no. 11, pp. 2508–2529, 2005.
[18] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[19] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[20] K. P. Bennett and E. J. Bredensteiner, "Duality and geometry in SVM classifiers," in *Proc. Int. Conf. Machine Learning*, 2000, pp. 57–64.
[21] D. Crisp and C. Burges, "A geometric interpretation of nu-SVM classifiers," *Adv. Neural Inform. Processing Syst.*, vol. 12, pp. 244–250, 2000.
[22] R. Herbrich, *Learning Kernel Classifiers: Theory and Algorithms*. Cambridge, MA: MIT Press, 2002.

Kazushi Ikeda (M'94) was born in Shizuoka, Japan, in 1966. He received the B.E., M.E., and Ph.D. degrees in mathematical engineering and information physics from The University of Tokyo, Tokyo, Japan, in 1989, 1991, and 1994, respectively.

From 1994 to 1998, he was with the Department of Electrical and Computer Engineering, Kanazawa University, Kanazawa, Japan. Since 1998, he has been with the Department of Systems Science, Graduate School of Informatics, Kyoto University, Kyoto, Japan. His research interests are focused on adaptive and learning systems, including neural networks, adaptive filters, and machine learning.